Part III — Unstructured Data

Data Retrieval:

III.1 Unstructured data and data retrieval

Statistical Analysis of Data:

III.2 Data scales and summary statistics

III.3 Hypothesis testing and correlation

III.4 χ^2 and collocations

Lecture timetable

Tuesday	8 March	Lecture	Tutorial 7, Information Retrieval
Friday	11 March	No Lecture	Assignment due to ITO 4pm
Tuesday	15 March	Lecture	Tutorial 8, Statistical Analysis
Friday	18 March	No Lecture	
Tuesday	22 March	Lecture	Tutorial 9, assignment feedback

Analysis of data

There are many reasons to *analyse* data.

Two common goals of analysis:

- Discover implicit structure in the data.

  E.g., find patterns in empirical data (such as experimental data).

- Confirm or refute a hypothesis about the data.

  E.g., confirm or refute an experimental hypothesis.

Statistics provides a powerful and ubiquitous toolkit for performing such analyses.

Data scales

The type of analysis performed (obviously) depends on:

- The reason for wishing to carry out the analysis.
- The type of data to hand: for example, the data may be *quantitative* (i.e., numerical), or it may be *qualitative* (i.e., descriptive).

One important aspect of the kind of data is the form of *data scale* it belongs to:

- *Categorical* (also called *nominal*) and *Ordinal* scales (for qualitative data).
- *Interval and ratio* scales (for quantitative data).

This affects the ways in which we can manipulate data.

Categorical scales

Data belongs to a *categorical scale* if each *datum* (i.e., data item) is classified as belonging to one of a fixed number categories.

Example: The British Government might classify Visa applications according to the nationality of the applicant. This classification is a categorical scale: the categories are the different possible nationalities.

Example: Insurance companies classify some insurance applications (e.g., home, possessions, car) according to the postcode of the applicant (since different postcodes have different risk assessments).

Categorical scales are sometimes called *nominal scales*, especially in cases in which the value of a datum is a name.

Ordinal scales

Data belongs to an *ordinal scale* if it has an associated ordering but arithmetic transformations on the data are not meaningful.

Example: The *Beaufort wind force scale* classifies wind speeds on a scale from $\mathbf{0}$ (calm) to $\mathbf{12}$ (hurricane). This has an obvious associated ordering, but it does not make sense to perform arithmetic operations on this scale. E.g., it does not make much sense to say that scale $\mathbf{6}$ (strong breeze) is the average of calm and hurricane force.

Example: In many institutions, exam results are recorded as grades (e.g., A,B,. . . , G) rather than as marks. Again the ordering is clear, but one does not perform arithmetic operations on the scale.

Interval scales

An *interval scale* is a numerical scale (usually with real number values) in which we are interested in *relative value* rather than *absolute value*.

Example: Points in time are given relative to an arbitrarily chosen zero point. We can make sense of comparisons such as: moment x is 2009 years later than moment y. But it does not make sense to say: moment x is twice as large as moment z.

Mathematically, interval scales support the operations of subtraction (returning a real number for this) and weighted average.

Interval scales do not support the operations of addition and multiplication.

Ratio scales

A *ratio scale* is a numerical scale (again usually with real number values) in which there is a notion of *absolute value*.

Example: Most physical quantities such as mass, energy and length are measured on ratio scales. So is temperature if measured in kelvins (i.e. relative to absolute zero).

Like interval scales, ratio scales support the operations of subtraction and weighted average. They also support the operations of addition and of multiplication by a real number.

Question for physics students: Is time a ratio scale if one uses the Big Bang as its zero point?

Visualising data

It is often helpful to *visualise* data by drawing a *chart* or plotting a *graph* of the data.

Visualisations can help us guess properties of the data, whose existence we can then explore mathematically using statistical tools.

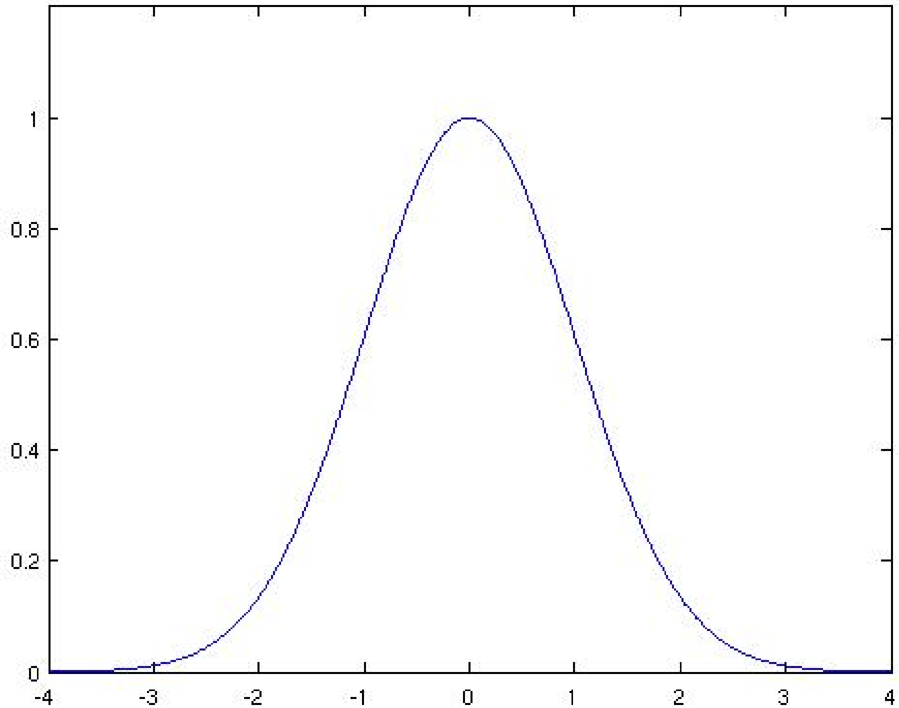
For a collection of data of a categorical or ordinal scale, a natural visual representation is a *histogram* (or *bar chart*), which, for each category, displays the number of occurrences of the category in the data.

For a collection of data from an interval or ratio scale, one plots a *graph* with the data scale as the x-axis and the frequency as the y-axis.

It is very common for such a graph to take a bell-shaped appearance.

Normal distribution

In a *normal distribution*, the data is clustered symmetrically around a central value (zero in the graph below), and takes the bell-shaped appearance below.



Normal distribution (continued)

There are two crucial values associated with the normal distribution.

The *mean*, μ, is the central value around which the data is clustered. In the example, we have $\mu = 0$.

The *standard deviation*, σ, is the distance from the mean to the point at which the curve changes from being *convex* to being *concave*. In the example, we have $\sigma = 1$. The larger the standard deviation, the larger the *spread* of data.

The general equation for a normal distribution is

$$y = c\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

(You do not need to remember this formula.)

Statistic(s)

A *statistic* is a (usually numerical) value that captures some property of data.

For example, the mean of a normal distribution is a statistic that captures the value around which the data is clustered.

Similarly, the standard deviation of a normal distribution is a statistic that captures the degree of spread of the data around its mean.

The notion of *mean* and *standard deviation* generalise to data that is not normally distributed.

There are also other, *mode* and *median*, which are alternatives to the mean for capturing the “focal point” of data.

Mode

Summary statistics summarise a property of a data set in a single value.

Given data values $x_1, x_2, \ldots, x_N$, the *mode* (or *modes*) is the value (or values) x that occurs most often in $x_1, x_2, \ldots, x_N$.

Example: Given data: $6, 2, 3, 6, 1, 5, 1, 7, 2, 5, 6$, the mode is 6, which is the only value to occur three times.

The mode makes sense for all types of data scale. However, it is not particularly informative for real-number-valued quantitative data, where it is unlikely for the same data value to occur more than once.
(This is an instance of a more general phenomenon. In many circumstances, it is neither useful nor meaningful to compare real-number values for equality.)

Median

Given data values $x_1, x_2, \ldots, x_N$, written in non-decreasing order, the *median* is the middle value $x_{(\frac{N+1}{2})}$ assuming N is odd. If N is even, then any data value between $x_{(\frac{N}{2})}$ and $x_{(\frac{N}{2}+1)}$ inclusive is a possible *median*.

Example: Given data: $6, 2, 3, 6, 1, 5, 1, 7, 2, 5, 6$, we write this in non-decreasing order:

$$1, 1, 2, 2, 3, 5, 5, 6, 6, 6, 7$$

The middle value is the sixth value 5.

The median makes sense for ordinal data and for interval and ratio data. It does not make sense for categorical data, because categorical data has no associated order.

Mean

Given data values $x_1, x_2, \ldots, x_N$, the *mean* μ is the value:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Example: Given data: $6, 2, 3, 6, 1, 5, 1, 7, 2, 5, 6$, the mean is

$$\frac{6+2+3+6+1+5+1+7+2+5+6}{11} = 4.$$

Although the formula for the mean involves a sum, the mean makes sense for both interval and ratio scales. The reason it makes sense for data on an interval scale is that interval scales support *weighted averages*, and a mean is simply an equally-weighted average (all weights are set as $\frac{1}{N}$).

The mean does *not* make sense for categorical and ordinal data.

Variance and standard deviation

Given data values $x_1, x_2, \ldots, x_N$, with mean μ, the *variance*, written $\mathbf{Var}$ or σ^2, is the value:

$$\mathbf{Var} \;=\; \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

The *standard deviation*, written σ, is defined by:

$$\sigma \;=\; \sqrt{\mathbf{Var}} \;=\; \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Like the mean, the standard deviation makes sense for both interval and ratio data. (The values that are squared are real numbers, so, even with interval data, there is no issue about performing the multiplication.)

Variance and standard deviation (example)

Given data: $6, 2, 3, 6, 1, 5, 1, 7, 2, 5, 6$, we have $\mu = 4$.

$$\begin{aligned} \mathrm{Var} &= \frac{2^2+2^2+1^2+2^2+3^2+1^2+3^2+3^2+2^2+1^2+2^2}{11} \\ &= \frac{4+4+1+4+9+1+9+9+4+1+4}{11} \\ &= \frac{50}{11} \\ &= 4.55 \quad \text{(to 2 decimal places)} \end{aligned}$$

$$\begin{aligned} \sigma &= \sqrt{\frac{50}{11}} \\ &= 2.13 \quad \text{(to 2 decimal places)} \end{aligned}$$

Populations and samples

The discussion of statistics so far has been all about computing various statistics for a given set of data.

Very often, however, one is interested in knowing the value of the statistic for a whole *population* from which our data is just a *sample*.

Examples:

- Experiments in social sciences where one wants to discover some general property of a section of the population (e.g., teenagers).
- Surveys (e.g., marketing surveys, opinion polls, etc.).
- In software design, understanding requirements of users, based on questioning a sample of potential users.

In such cases it is totally impracticable to obtain exhaustive data about the population as a whole. So we are forced to obtain data about a sample.

Sampling

There are important guidelines to follow in choosing a sample from a population.

- The sample should be chosen *randomly* from the population.
- The sample should be as *large* as is practically possible (given constraints on gathering data, storing data and calculating with data).

These two guidelines are designed to improve the likelihood that the sample is *representative* of the population. In particular, they minimise the chance of accidentally building a *bias* into the sample.

Given a sample, one calculates statistical properties of the sample, and uses these to infer likely statistical properties of the whole population.

Important topics in statistics (beyond the scope of D&A) are *maximising* and *quantifying* the reliability of such techniques.

Estimating statistics for a population given a sample

Typically one has a (hopefully representative) sample $x_1, \ldots, x_n$ from a population of size N where $n \ll N$ (i.e., n is much smaller that N).

We use the sample $x_1, \ldots, x_n$ to estimate statistical values for the whole population.

Sometimes the calculation is the expected one, sometimes it isn't.

The best estimate m of the *mean* μ of the population is:

$$m = \frac{\sum_{i=1}^{n} x_i}{n}$$

As expected, this is just the mean of the sample.

Estimating variance and standard deviation of population

To estimate the *variance* of the population, calculate

$$\frac{\sum_{i=1}^{n}(x_i - m)^2}{n-1}$$

The best estimate s of the *standard deviation* σ of the population, is:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - m)^2}{n-1}}$$

N.B. These values are *not* simply the variance and standard deviation of the sample. In both cases, the expected denominator of n has been replaced by $n-1$. This gives a better estimate in general when $n \ll N$.

Caution

The use of samples to estimate statistics of populations is so common that the formula on the previous slide is very often the one needed when calculating standard deviations.

Its usage is so widespread that sometimes it is wrongly given as the definition of standard deviation.

The existence of two different formulas for calculating the standard deviation in different circumstances can lead to confusion. So one needs to take care.

Sometimes calculators make both formulas available via two buttons: σ_n for the formula with denominator n; and σ_{n-1} for the formula with denominator $n - 1$.

Further reading

There are many, many, many books on statistics. Two very gentle books, intended mainly for social science students, are:

P. Hinton	First Steps in Statistics
Statistics Explained	D. B. Wright
Routledge, London, 1995	SAGE publications, 2002

These are good for the formula-shy reader.

Two entertaining books (the first a classic, the second rather recent), full of examples of how statistics are often misused in practice, are:

D. Huff	M. Blastland and A. Dilnot
How to Lie with Statistics	The Tiger That Isn't
Victor Gollancz, 1954	Profile Books, 2008